DOCUMENT RESUME

ED 298 264 CE 050 757

AUTHOR Martin, Randy
TITLE Developing and Improving the Quality of Written Tests.
PUB DATE 22 Sep 88
NOTE 9p.; Paper presented at the Midwest Nuclear Training Association's Annual Nuclear Instructors' Workshop (3rd, Columbus, OH, September 22, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Item Analysis; Item Banks; Postsecondary Education; *Test Construction; Test Format; Test Interpretation; *Test Reliability; *Test Validity; Vocational Education

ABSTRACT

Reasons for administering tests fall into two categories--decision-making and promoting learning. The two bases of tests are learning objectives and the level of learning at which training is developed. Test development involves a number of steps. The best way to tie objectives to test items is through the use of a table of specifications, which organizes test development, specifies level of learning, determines relative importance of subject matter, and promotes uniformity across tests. Another part of test planning is the determination of test item formats to be used. Test item development follows these steps: write, review, rewrite, pilot. After test items are developed, they should be put into a test item bank. The next step is preparation for the organization and administration of the examination. Written tests can be improved only one way--by increasing the reliability and validity of the test. One activity that is very beneficial is the process of test item analysis. It includes three major activities--item difficulty, item discrimination, and alternative analysis. A test is a measuring device, and the numbers must be interpreted to have meaning as an evaluation. Part of the interpretation of test data is the application of an appropriate reference. Three major reference group classifications are content-referenced, group-referenced, and criterion-referenced. (YLB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298264

DEVELOPING AND IMPROVING THE QUALITY OF WRITTEN TESTS

PRESENTED BY RANDY MARTIN

INPO

AT

MIDWEST NUCLEAR TRAINING ASSOCIATION

THIRD ANNUAL NUCLEAR INSTRUCTOR'S WORKSHOP

THURSDAY, SEPTEMBER 22, 1988

CE 050 757

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R Martin

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Introduction

In this session I would like to discuss those steps essential to the development of high quality written examinations. It is important that all test developers and instructors have an awareness of these steps. The topics we will discuss are:

1. Purpose of tests
2. Bases of tests
3. Developing a test
4. Test improvement
5. Test interpretation

The Role of Tests

There are several reasons for administering tests, but they generally fall into one of two categories, decision-making tools or learning tools. We generally expect tests to tell us what a person can do or what he is capable of doing, or how well he can do it. The test gives us a measure of the trainee's skill or ability which, with an appropriate reference, can be used to make decisions.

A second purpose of testing is to promote learning. Tests are used to motivate people, teach material, or reinforce previously learned material. The pop quiz is as valuable in motivating students as it is in evaluating progress. Reviewing test results with trainees can correct misconceptions and reinforce learning. While our main interest is in using tests for evaluation, we should not neglect their value as learning tools.

Bases of the Test

The systematic development of training requires that objectives be established for instructional activities. These same objectives must be used for exam development.

An important aspect of a learning objective is the level of learning at which training is developed. We teach basic facts, the ability to apply knowledge, and the skills of analysis. Each of these purposes requires different teaching techniques. To measure mastery of these skills requires different test questions. The level of learning must be established as part of the learning objective and this level of learning must be accurately tested. Several theoretical models have been created to aid in this process. The most commonly used is that developed by Benjamin Bloom. His hierarchy specifies five levels of learning: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. My 7th grade son is currently learning Bloom's taxonomy. He asked me why he had to learn it! I didn't have an answer. I'm not sure why its important to 7th graders. But I do know why its important to instructors and course developers. Instruction must be developed at the appropriate level. Tests must be developed to match

that level, and the best way to recognize and assure that this occurs is through the understanding and application of a learning hierarchy.

Where does one find instruction and procedures for test development? There are literally hundreds of published books and articles on this topic. The 1988 Guide to Books in Print lists 191 publications on the topic of educational tests and measurement. INPO has three publications that address test and test item development. INPO 85-006, Principles of Training System Development, and two addendums, INPO 88-001(Addendum 1), Test Item Development, and INPO 88-001(Addendum 2), Examinations: Design, Development, and Implementation. Because of the recent interest in developing open book exams, NUMARC has drafted guidance for this type of examination process. Educational Testing Service (ETS) produces several test and measurement publications. Their "Tests and Measurement Kit" contains four booklets that clearly illustrate the development, use, and evaluation of teacher-made tests.

Test Development

I'd like to shift now to some considerations for the development of effective tests. A good test begins with planning. Our first question might logically be - What is the purpose of the test? It may be an informal quiz primarily intended to motivate or instruct, but most often it will be a formal exam intended to evaluate some major aspect of trainee performance. The rest of my comments today will address this formal testing for evaluation purposes. We've already discussed objectives, the need for them in training development, and the necessity of tying test items to objectives. One of the best ways to make this connection is through the use of a table of specifications. The table of specifications is effective in accomplishing several things. It can be used to:

1. Organize test development
2. Specify level of learning
3. Determine relative importance of subject matter
4. Promote uniformity across tests

Once developed, the table of specifications should be made known to students, trainers, and program evaluators.

Another part of test planning is the determination of test item formats to be used. The format for a question must be matched to the objective being tested. It is more difficult to develop test questions that address higher level learning objectives. Not all item formats can satisfactorily address all learning levels. Test item formats can be grouped into the following categories:

1. Oral -- not generally used in formal evaluation, but effective for classroom review of material.
2. Structured -- examples are T/F, multiple choice.

3. Unstructured -- an essay question is an unstructured item. It allows considerable latitude in trainee response. It is ideal when evaluating creativity, organization of thought, or ability to analyze.

4. Performance -- actually performing some task. A JPM is an example of a performance item.

With the planning complete, the test item can be developed. There are many things to know about the development of test items. Many of these are incorporated into rules of do's and don'ts. Avoid double negatives. Make all distracters plausable. Avoid the use of specific determiners (always, never). Do not include more than a single idea in a true-false item. Unfortunately, no matter how long the list, these rules cannot replace a skilled test developer. Test writing skills are developed over time. A test developer must have a thorough knowledge of the subject matter, be a logical thinker, and a master of verbal communication.

Test item development must be logical and repetative. **Write. Review. Rewrite. Pilot.** (WRRP) This is a simplification, but reflects the necessary steps of the development process.

After you have developed a good test item and used it in an exam, don't throw it away. Put it in the bank. A test item bank. Saving test questions is like saving dollars. They are handy to have in the future. While test item banks are good to have, they have certain requirements.

1. Security -- make sure only approved personnel have access.

2. Guidelines for use -- determine how the bank will be administered. What it will contain? Establish a system for use.

3. Numbering scheme -- this is necessary for access. Be sure you can access data several ways, by topic, objective, table of specifications, question number, etcetra.

If possible include provisions for test item analysis and the selection of test items based on the table of specifications.

Having completed the development of test items, prepare for the organization and administration of the exam.

1. Develop a clear logical format for the exam based on the table of specifications.

2. Provide test directions, both written and oral.

3. Monitor the exam to prevent compromise and to insure proper understanding.

4. Accurately score the exam.

A recent interest has developed in the use of references during the examination process. This is based, in part, on the recent development by the NRC of open

reference exams. Open reference exams are not that different from regular exams. However some points should be stressed when using references (open book).

1. Separate the testing of required knowledge from the testing of reference usage. Test closed book items separately and before open book items. Frequently the reference material provides answers to the knowledge (closed book) questions.

2. Testing should be for the proper use and interpretation of the references.

3. Insure that the trainee is required to apply his knowledge of the references; don't use a test item that has already been used in class, this item can often be answered by recall rather than application.

4. The use of references adds reality to the testing situation.

Improving the Quality of Written Tests

We've talked of developing an exam, but once developed how do we improve its quality? There is just one way. We must increase the reliability and validity of the test. To do this we must first understand some basic concepts.

Reliability means consistency. A reliable car travels 100,000 miles without major repair. A reliable test measures the same thing each time it is administered. Test reliability is measured by readministering the test and then comparing the results. Test reliability is composed of both internal and external factors. Many factors affect the measured reliability of a test.

To increase test reliability:

1. Increase test length

2. Increase homogeneity of test items

3. Make test items more discriminating

4. Use items of medium difficulty

5. Administer to a wider range of examinees

To determine test reliability statistical tools are used. Most of the commonly used statistical tools are designed for norm-referenced testing, but they can also be used for criterion-referenced tests. The evidence does not support the need to use separate criterion-referenced reliability formulas.

When considering the reliability of exams a related issue should be considered. That is the issue of test accuracy for individual trainees. Test reliability is calculated from group data, and the results only apply to the group. In many, if not most situations, we want to interpret individual scores. To do this we need to understand error of measurement concepts. The statistic known as the standard error of measurement (SEM) shows the fallacy of accepting any single exam score as being an accurate score for an individual. Due to chance and the inaccuracy of any exam, measured scores are only approximate indicators of true scores.

Instructors need to be aware of this and not put excessive confidence in individual scores.

Validity. A test that is valid is a test that measures what it is supposed to measure. Validity is the most important requirement of a test. A test that is not valid is worthless. While a test can be reliable and not valid, it is impossible for a test to be valid and not reliable. A test should be both. Validity of a test has to do with the test itself and the proper usage of the test. If the usage is not appropriate the test results will not be valid for the intended purpose. Validity cannot be measured directly. Instead we must collect evidence that a test is valid. There are three major categories of validity evidence.

1. Content -- use subject matter experts to review the test content. Content should accurately reflect the content taught and the stated objectives.

2. Criterion -- a suitable criteria is determined, and the test results are compared to this criterion. Performance on the job is the most desired criterion for many training exams, but it is hard to quantify. This type of validity is called predictive validity if the criterion selected occurs at a later point in time.

3. Construct -- refers to the underlying theory of the exam. I.Q. (intelligence quotient) is a construct. How well does the test measure the construct? For most practical exams the constructs are obvious. We want a math test to measure math skills, and a welding performance should measure welding skill.

Now that we have an understanding of these basic testing concepts, how do we actually go about the process of increasing the validity and reliability of our exams? Actually we've already discussed many of these important activities.

1. Table of Specifications

2. Content review by subject matter experts

3. Review of objectives

4. Hierarchy of learning objectives

Another activity that is very beneficial, but seldom used is the process of test item analysis. Test item analysis includes three major activities.

1. Item difficulty -- is the test item too easy or too hard? Medium difficulty items discriminate best and are recommended for norm-referenced test situations. For criterion-referenced exams many test items will be answered correctly by most or all trainees. For these exams item difficulty is primarily used to indicate group progress on given objectives.

2. Item discrimination -- a good test item should exhibit positive discrimination values. This means that more knowledgeable trainees answer the test item right more often than less knowledgeable trainees.

3. Alternative analysis -- looks at the test item alternatives. Which of the distracters are effective foils, and which do not work. Many test items have distracters that are non-functional and should be replaced.

The use of item analysis procedures is an excellent method for improving exams. However do not let the process become a controlling factor in test development. The process is excellent for fine-tuning an exam. But a test item, no matter how good it may be technically, is worthless if it doesn't match the objective it is supposed to test. This brings us to the next topic I would like to discuss.

Interpreting Test Scores

It should be kept in mind that a test is a measuring device; like a ruler, the numbers must be interpreted to have meaning as an evaluation. Part of the interpretation of test data is the application of an appropriate reference. I talked earlier of norm-referenced and criterion-referenced testing. I would like to be more specific now and discuss three major reference group classifications. They are:

1. Content-referenced

    a. objectives-referenced -- have the objectives been met? What skills have been demonstrated?

    b. domain-referenced -- when not all objectives are tested, but only a sample is selected from the domain of interest.

2. Group-referenced

    a. norm-referenced -- scores are compared to a group score.

    b. treatment -- scores are compared to treatment groups, controlled versus noncontrolled group, treatment group A, treatment group B, etcetra.

3. Criterion-referenced

Criterion-referenced tests establish a cut-off score. A trainee either meets or fails to meet the cut-off. The decision is go/no go, pass/fail. However the basis on which the cut-off score is established can be content- or norm-referenced.

    a. Content-referenced -- the reference for the cut-off is how much content knowledge the trainee has acquired.

    b. Norm-referenced -- the cut-off score is referenced to a norm group. The better the norm group does, the higher the cut-off score will be.

As you may suspect there is a major problem with criterion-referenced test interpretation. We can easily determine who passes or fails after the criterion cut-off score is established, but how are we to know where to set the criterion? What is an acceptable level of performance? Many training programs use a content referenced cut-off score for the criterion. Eighty percent is common. But is this level valid?

A method to establish cut-off score validity is to have subject matter experts (SMEs) examine each test item. From this it can be determined which item, or item parts must be answered correctly for a trainee to be considered acceptable on the job. Summing all the required knowledge identifies the acceptable level of Performance (APL).

Summary

To summarize our discussion this morning, we have overviewed issues important in the development and improvement of examinations. They were:

1. Purpose of Tests
   a. decision-making
   b. promote learning
2. Bases of Tests
   a. learning objectives
   b. level of learning
3. Developing a Test
   a. table of specifications
   b. item development
   c. test bank
   d. test administration
4. Test Improvement
   a. increase validity and reliability
   b. use item analysis
5. Test Interpretation
   a. content-referenced
   b. group-referenced
   c. criterion-referenced